The wild wild web: anonymity and racial prejudice in online culture

by

Tiffany Jo Lawless

B.S., Cornell College, 2015

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Donald Saucier

Copyright

Abstract

Research has shown people view some anonymous communication as less honest (e.g., Bergstrom, 2011), but some as more honest (e.g., Sticca & Perren, 2012), than identifiable communication. This discrepancy may be due to whether the target of a post is a group of people or an individual. Across two studies, I examined the effects of specificity of target on perceptions of honesty in prejudiced speech by manipulating whether posts appeared to be posted anonymously/identifiably and whether the content targeted a racial group/named individual in a counterbalanced within-groups design. In Study 2, I also manipulated whether posts were public/private messages. Generally, posts targeting individuals were rated as equally racist, but more honest, than posts targeting groups. Additionally, public anonymous posts were rated as less honest than other posts. These studies imply people may disregard anonymous expressions of prejudice, dismissing them as dishonest. These studies demonstrate that many people may not take anonymous online prejudiced rhetoric seriously, which could foster toxic online environments conducive to incitements of real-world violence against marginalized groups.

Table of Contents

List of Tables

Acknowledgements

I would like to thank my committee members, Donald Saucier, Gary Brase, and Laura Brannon as well as my friends and family, with special thanks to Stacy Stoffregen, Angela Rose, Frank Giordano, Nathaniel Voss, Kelsey Couture, Kevin Kenny, Conor O'Dea, Stuart Miller, Amanda Martens, Svyataslav Prokhorets, Evelyn Stratmoen, and Richard Reed.

Dedication

This work is dedicated to wholesome Redditors and other decent denizens of the interweb.

Chapter 1 - The Wild Wild Web

The recent rise of online social media has opened easily accessible routes to anonymity for many. Although this has allowed some people to more freely share stories and productive discourse, it has also created new avenues for the expression of hate and prejudice. When people behave anonymously, they are less likely to display altruism (Locey & Rachlin, 2015), and more likely to exhibit antisocial behavior (Nogami & Takai, 2008). Anonymous posts on online forums tend to be more aggressive than identifiable posts (Moore, Nakano, Enomoto, & Suda, 2012), and young people who believe they are behaving anonymously display cyber aggression more often (Wright, 2013). The tendency of anonymity to negatively impact behavior is widely documented and is generally accepted by both the scientific and lay communities (Reader, 2012).

Anonymity is the degree to which the receiver or observer of communication perceives the communication's source as unknown or unspecified (Scott, 1998). Anonymity has become an integral facet of online communication, with the advent of anonymous usernames, "throwaway" accounts, and the ability for one person to hold and automatically communicate from hundreds of email addresses and social media accounts at one time. Whereas chatting anonymously online with others had comparatively innocent origins in cooperative text-based videogames and fandom-based message boards, it has now become a tactic employed by both widely political 'hacktivism' groups such as Anonymous as well as seemingly innocuous, but arguably insidious, "trolls" and cyberbullies (Crawford, 2009; Wang, Wang, Wang, Nika, Zheng, & Zhao, 2014). Removing traditional concepts of strong identities and social links, largely anonymous social media platforms, such as Reddit, Whisper, Discord, 4chan, and formerly Yik Yak, encourage communication between strangers and allow users to express themselves without fear of real-life consequences (Crawford, 2009; Wang, et al., 2014).

As opportunities for online communication have grown, there has also been an increase in new social problems, such as cyberbullying, particularly among adolescents. Hinduja and Patchin (2006) define cyberbullying as a deliberate, repeated, and hurtful activity using a computer, mobile phone, or other electronic device. The rapid growth in technology presents many avenues for cyberbullies to engage in negative and hurtful behavior, and it also allows cyberbullies to remain unseen if they choose an anonymous platform. The cyberbully can quickly use a digital device to post or send a hurtful message to a larger group of people while remaining unseen. By using Facebook, someone who cyberbullies can post a hurtful message about his, her, or their targets, and within minutes, this is broadcast into cyberspace to the target's real-life friends and acquaintances. A hurtful message will be seen in seconds by hundreds of online users. Even when deleted, a typed message can be re-discovered because it is never completely removed from the Internet. According to Wong-Lo and Bullock (2011), in 2010, incidences of cyberbullying increased to 90% of youth playing one of the three roles—third-party observer, bully, or target—even though only 19% of youth had played one of these roles in cyberbullying in 2000 (Ybarra & Mitchell, 2004). One way cyberbullying is happening within the digital realm is on social networking sites such as Facebook. Using Facebook and other social media platforms as a means to engage in cyberbullying is happening almost universally among teens; 95% of teens on social media have witnessed cruel behavior on social media sites (Lenhart, Madden, Smith, Purcell, Zickuhr, & Rainie, 2011).

Along with identifying the prevalence of cyberbullying, early research created profiles of those involved in cyberbullying as targets and/or bullies. Troublingly, one study found people with special needs, unusual academic abilities, poor social skills, odd or undesirable physical appearances, physical and mental disabilities, unfashionable clothing, and those of a minority

ethnicity were often targeted (Cassidy, Jackson, & Brown, 2009). Thus, seeing racist and other prejudiced attacks online is a common situation. This may desensitize observers to this kind of communication, and this may result from and help create an online culture that is socially and morally distinct from face-to-face culture.

Runions (2015) suggests that technological communication alters views towards targets; third-party observers experience moral disengagement from their normal values. The ambiguity created by technological communication alters perceptions of blame and empathy, with third-party observers justifying their reactions by perceiving the target as somehow to blame for the situation (Runions, 2015).

Additionally, third-party observers get involved in instances of cyberbullying at a greater rate compared to face-to-face bullying, and they directly alter the bullying experience through their involvement (Anderson, Bresnahan, & Musatics, 2014; Barlinska, Szuster, & Winiewski, 2013; Runions, 2015). The process of communicating through technology may alter the social context or way that third-party observers would interact with bullying targets (Calvete, Orue, Estévez, Villardón, & Padilla, 2010). Third-party observers may react to an environment that models aggression by contributing to the derogation of the targets. A second unique characteristic is that a target of cyberbullying may be unaware of who the bully is. One study found 48% of those bullied did not know who had bullied them because the bully had maintained an anonymous username or attacked via an anonymous platform (Kowalski, & Limber, 2007). One study found that when the cyberbullying was anonymous, there was little emotional impact to the target (Reeckman & Cannard, 2009). However, other studies had been unable to replicate this finding (e.g., Dilmac, 2009; Price & Dalgleish, 2010). Additionally, many findings have indicated a link between aggression and attention-seeking, and engaging in cyberbullying behaviors. (Harman,

Hansen, Cochran, & Lindsey, 2005; Li, 2005; Willard, 2007). Bullies could be engaging in cyberbullying for fun, to exert power, or both (Cassidy et al., 2009; Reeckman & Cannard, 2009). It is possible that observers understand the possibility of these motivations, and therefore attribute online prejudiced speech to honesty in some circumstances and to attention-seeking or "trolling" in others.

Hardaker (2010) defines a troll as a person "whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement" (pg. 237). Trolling permeates the online ecosystem. On some level, trolls are responsible for creating and/or amplifying many popular memes. LOLcats, RickRolling, the Guy Fawkes mask, advice animals, demotivational posters, and other attention-seeking gimmicks are common trolling behaviors. Trolls also frequently perform more antisocial and widely unacceptable behaviors, such as espousing racist or sexist rhetoric. Additionally, most trolls establish a number of firewalls between their online and offline personas, making their "true" intentions difficult if not impossible to discern, and carefully maintaining their anonymity (Bourdieu 2001; Dahlberg, 2001; Donath, 1998). Trolling is typically predicated on sensationalism, spectacle, and emotional exploitation, all of which can be achieved via extreme prejudiced rhetoric, particularly against large groups of people. Therefore, it is possible that people commonly perceive online anonymous prejudiced speech targeting groups of people as attention-seeking behavior rather than as honest dissemination of personal beliefs. As Donath (1998) argues, the existence of trolls, or even the possibility of the existence of trolls, makes community members less likely to trust outsiders and anonymous people as honest. This idea of trolling has permeated internet culture to the point that research has suggested that some communities (e.g., Reddit) use trolling accusations as social deterrents to prevent lying (Bergstrom, 2011). However, it is also

understood that engaging in racism or sexism or homophobia does not automatically make one a troll, and there is therefore general uncertainty regarding the honesty of anonymous people online.

On the other hand, there is evidence that the opportunity to communicate anonymously over the internet results in greater opportunities for honesty, and perhaps greater perceptions of honesty by third parties. Bargh, McKenna, and Fitzsimons (2002) found that certain social networking sites decrease barriers to communication, which then promotes greater self-disclosure. In an experiment involving undergraduate students who were randomly assigned to interact with conversation partners in an online setting or in a face-to-face setting, those who were communicating online were better able to express their 'true-self' qualities (Bargh et al., 2002). Tidwell and Walther (2002) also found that participants who communicated via computer had a higher proportion of intimate and direct uncertainty reduction behaviors than those participants who met in face-to-face interactions.

Third-Party Perception

Whereas several studies have focused on how anonymity affects behavior, relatively little is known about how anonymity affects the perception of this behavior by third parties. Reader (2012) documented that both professional journalists and lay commenters on news websites viewed anonymous comments as more cowardly and less valuable to discussion than identifiable comments, even going so far as to dub such commenters "trolls" and to refuse to engage with them. These and other findings (e.g., Bergstrom, 2011; Reeckman & Cannard, 2009) suggest that people might generally view anonymous online communication as worthless noise rather than as honest expressions of opinions. However, Sticca and Perren (2012) found that adolescents view anonymous cyberbullying as more severe and painful than identifiable cyberbullying. These and

other findings (e.g., Bargh et al., 2002; Tidwell & Walther, 2002) suggest that people may generally view anonymous online communication as more honest than identifiable communication. This discrepancy could be the result of people attributing different types of hateful behavior to different purposes when the perpetrator is anonymous compared to when the perpetrator's identity is known. Thus, the current research question is: do people attribute online prejudiced speech to different purposes when the perpetrator is anonymous compared to when the perpetrator's identity is known?

Toxic Online Disinhibition

Hatred directed at members of groups due to factors they cannot control, such as their race, is not new, but it has taken on a new dimension in the online world. Online hate involves actions such as the denigration, harassment, and exclusion of, as well as the advocacy of violence toward, specific groups (Hawdon, Oksanen, & Räsänen, 2017; Räsänen, Hawdon, Holkeri, Keipi, Näsi, & Oksanen, 2016; Sponholz, 2018). The online environment involves anonymity, invisibility, asynchronicity, textuality, and lack of face-to-face contact, and punishment and repercussions are considered less likely to occur as compared with the offline world (Suler, 2004). These circumstances can promote rude language, hatred, and threats. This promotional tendency is also referred to as toxic online disinhibition (Suler, 2004). Toxic online disinhibition can also decrease the ability for empathy, self-control, and to recognize social cues (Suler, 2004; Voggeser, Singh, & Göritz, 2018). When compared to the offline world, there is an increased likelihood that fewer adults are present in the online world of adolescents, which can also increase aggressive behavior and discrimination (Hinduja & Patchin, 2008; Tynes, Reynolds, & Greenfield, 2004). Past research has revealed that higher levels of toxic online disinhibition are positively associated with cyberbullying perpetration,

flaming, and trolling (Görzig & Ólafsson, 2013; Udris, 2014; Voggeser et al., 2018; Wright, 2013; Wright, Harper, & Wachs, 2018). Therefore, it can be proposed that toxic online disinhibition might also lead to less self-monitoring when expressing beliefs through hateful or degrading speech online, making inappropriate attacks on minorities more likely.

The online disinhibition effect is a decrease in the reservedness of behavior frequently displayed in online environments (Joinson, 2007; Kiesler, Siegel, & McGuire, 1984; Suler 2004). Many behaviors that are performed online, particularly those performed anonymously, can be attributed to the online disinhibition effect (Joinson, 2001; Kiesler et al., 1984). These behaviors often manifest as overly aggressive and/or hateful posts or comments, and these hostile behaviors can be attributed to toxic online disinhibition (Suler, 2004). Posts attributable to toxic online disinhibition typically include aggressive language, swearing, and derogatory names (Dyer, Green, Pitts, & Millward, 1995). Such toxic and hostile behaviors can often be found, not only within hate-spewing blogs and instances of cyberbullying, but also in places as innocuous as online video gaming sites and the comments on YouTube videos (Chau & Xu, 2007; Huang & Chou, 2010; Moor, Heuvelman, & Verleur, 2010; Williams & Skoric, 2005). Given that anonymity is often a major factor in the development of toxic online disinhibition (Joinson, 2007; Lapidot-Lefler & Barak, 2012; Suler, 2004), and that people recognize this behavior as toxic and outside the general social norms (e.g., Reader, 2012), it is possible that people discount hateful and anonymous online activity and attribute less importance and truth to it.

Social Identity Model of Deindividuation Effects (SIDE) Theory

SIDE theory suggests that technological communication alters perceptions of oneself and others (Postmes, Spears, & Lea, 1998), contributing to conflict that promotes online prejudice. Espousing prejudiced rhetoric is an inherently social process because of the number of third-

party observers that may witness the speech, particularly on the internet (Anderson et al., 2014; Barlinska et al., 2013). Via action or inaction, third-party observers can affect the severity of online prejudice for targets. Third-party observers commenting or forwarding a hateful message actively contribute to the bullying process, whereas third-party observers communicating support to the target may reduce the trauma associated with being targeted (Anderson et al., 2014). The application of SIDE to the issue of online expressions of prejudice examines how technological communication changes perceptions of identity, increasing the likelihood of acting in ways that differ from normal behavior (Postmes et al., 1998).

SIDE theory suggests that when individuals communicate through technology, a change in perception occurs (Postmes et al., 1998). Postmes and colleagues (1998) argue that the social definition participants give to a context affects how they communicate with each other through technology, and the features of that technology may in turn influence how the interaction unfolds. Postmes and Baym (2005) suggest that although the use of communication technology does not necessarily lead to uniform effects across situations, and technology does not determine interpersonal interactions en masse, it does have an influence on an individual and social level (Postmes & Baym, 2005). Namely, the features of technological communication highlight certain aspects of identity during online interactions, creating a shift in perceptions that can alter communication (Postmes et al., 1998). Technology leads certain elements of interactions to become more or less salient (Postmes et al., 1998). The noticeable effects of this salience are changes in perceptions of individual identity compared to social identity (Postmes, Spears, Sakhel, & de Groot, 2001). Postmes and Baym (2005) suggest that when communicating with others, individuals retain a sense of personal identity while maintaining a perception of social identity. SIDE postulates that the features of online communication heighten awareness of the

social context or group (Postmes & Baym, 2005). Accounting for why individuals place importance on social identity in the technological communication environment, Moral-Toranzo, Canto-Ortiz, and Gomez-Jacinto (2007) explain that it fulfills the need to belong and is tied to self-satisfaction. This heightened awareness of the group, or the disconfirming comments and lack of confirming comments to a target of prejudice could influence third-party observers' perceptions when considering whether and how they should respond to a hateful message. For example, individuals may engage in self-stereotyping, reinforcing their characteristics and opinions based on the predominant views of the group (Postmes & Spears, 2002). As the group identity becomes more salient, individuals are more likely to adhere to group norms. Additionally, individuals perceive a degree of anonymity when they communicate online, even though they know each other and interact in real life (Moral-Toranzo et al., 2007). Applied to third-party observers in prejudiced rhetoric, individuals might be more likely to comment in a certain way or avoid supportive actions towards targets than they would be within a real-life context because the established group norms indicate a culture where prejudice is to be expected and perhaps even condoned.

In addition, SIDE explains that when communicating through a technological context, perceptions of individual identities are reduced (Postmes & Baym, 2005). The process is called deindividuation, and can be used to explain why third-party observers might get directly involved in responding to hateful messages (Barlinska et al., 2013). Due to a shift in perception that occurs when communicating through technology, third-party observers may feel a need to respond in a way that reinforces social identity (Barlinska et al., 2013). Individuals consider how comments fit in with the social group viewing the message (Postmes & Baym, 2005), and lose

awareness of comments being directly received by the target, with a lack of understanding of how the target is adversely affected.

Deindividuation occurs when individuals experience reduced awareness of themselves and of others (e.g., Carr, Vitak, & McLaughlin, 2011; Postmes et al., 1998). SIDE suggests that anonymity is a key factor in determining how deindividuation occurs (Postmes et al., 2001). Postmes and colleagues (1998) argue that the way communication unfolds through technology can lead to a change in cognitive processing. A typical response to an interpersonal situation changes as anonymity reduces perceptions of personal identity and magnifies views of group identity. Postmes and Baym (2005) suggest that, in social interactions, individuals have a sense of both individual and social identities, but group membership is often exaggerated in an online setting (Walther & Bazarova, 2007). The asynchrony of communication and the unique ability of several others to respond to a social media message facilitates the perception of communicating with a group even if a message is directed to one member (Postmes & Baym, 2005). As a result, when a third-party observer views a hateful message, the communication may be considered as reflective of the group, rather than personal communication. A third-party observer that experiences deindividuation would pay more attention to the social context, or the comments of others, rather than considering how the response, or lack of a supportive response, directly impacts the target. Deindividuation increases in situations with greater anonymity (Postmes et al., 2001). A lack of distinguishing features alters perceptions of the self and of others as individuals (Postmes & Baym, 2005). Illustrating this point, Postmes and Spears (2002) found increasing anonymity by manipulating perceptions of personal identity led to a greater use of gendered stereotypes, and participants exhibited little concern that the comments would trace back to them.

Additionally, deindividuation leads to a perception of a breakdown of traditional social barriers (e.g., Postmes et al., 2001). As a result, individuals may feel emboldened in their actions, behaving without inhibitions or communicating antisocially, and thus behaving differently from how they normally would in a real-life context (e.g., Postmes et al., 2001). Situations characterized by anonymity appear to change social barriers by facilitating more negative behaviors, instead of promoting equality. In the case of online prejudice, third-party observers may feel emboldened to like, share, or leave disparaging messages for targets. For example, Slonje, Smith, and Frisen (2012) applied the concept of deindividuation to explain why adolescents would act as cyberbullies. The authors suggested that deindividuated cyberbullies would feel less guilt and remorse for actions, because the perception of directly bullying another individual is reduced in a cyber context. Similarly, third-party observers may feel a lack of remorse about disparaging actions or inaction to support cyberbullying targets as a result of being deindividuated. Anonymity in cyberbullying reduces pressure and constraints when communicating with targets (Calvete et al., 2010). When a social media platform alters social cues and offers a sense of protection through anonymity, internet users may feel emboldened similar to the findings of anonymity in situations of mob mentality (Calvete et al., 2010; Runions, 2013). Barlinska and colleagues (2013) suggest that the sense of deindividuation is furthered by a lack of direct feedback from targets. In other words, without targets articulating the harm they experienced, the sense of one's actions negatively affecting the targets is lessened. Thus, internet users experience reduced responsibility for behaviors. Anonymity is an influential aspect accounting for cyberbullying, and SIDE provides a framework showing the effects of anonymity on the process of deindividuation. Another aspect of deindividuation should be

considered when examining cyberbullying, and that is the way individuals communicate out of consideration to group norms.

SIDE theory is a reinvention of classic deindividuation theory that places more emphasis on situational circumstances in social contexts (e.g., Christopherson, 2007). SIDE theory postulates that, when all group members are anonymous, group salience and member identification with the group both increase (e.g., Spears & Lea, 1992). However, if some group members are identifiable while some are anonymous, SIDE theory postulates that the anonymous members will identify less with the group and more with themselves. Therefore, anonymous members are more likely to behave in ways that are detrimental to the group (Spears & Lea, 1992). This includes prejudiced rhetoric and aggression, assuming that the group does not promote such conduct. This explains why certain websites, such as YouTube or news outlets, where commenters can choose to be anonymous or identified, elicit more hateful anonymous behavior than websites such as Facebook or Whisper, where virtually all members are identifiable or all members are anonymous. The common association of anonymous commenters with meaningless prejudiced rhetoric and detrimental actions might suggest internet users generally attribute anonymous posts to lower levels of honesty and higher levels of attention seeking. However, the type of prejudiced rhetoric or detrimental action as well as the individual differences and biases present in third-party observers (e.g., social desirability, racial attitudes) may affect the degree to which those observers attribute anonymous posts to honesty and attention-seeking.

Social Desirability

The validity and value of psychological assessment rests, to a large extent, on accurate responding. Over- or underreporting of perceptions, thoughts, and feelings can easily invalidate

the results of psychological assessments by contributing faulty data. As a result, intentional misreporting has represented a significant and controversial concern to the field for decades (Ben-Porath & Waller, 1992; Cashel, Rogers, Sewell, & Martin-Cannici, 1995; Nichols & Greene, 1997). Due to the sensitive nature of prejudice, some individuals may be motivated to respond in a manner that makes them look non-racist or otherwise "good." For example, it has been found that people may wish to present a more favorable impression of themselves, such as by endorsing positive traits (Bagby et al., 1999). Research has convincingly shown that psychological measures can be positively distorted across a broad spectrum of settings, from job applications to inpatient units, often while successfully avoiding detection (e.g. Baer & Miller, 2002; Bagby et al., 1999; Pauls & Crost, 2005; Viswesvaran & Ones, 1999). Additionally, individuals may be reluctant to disclose undesirable personality traits, such as racism, because of the social pressures they have experienced in the past (Martin, Pescosolido, & Tuch, 2000). Social desirability likely motivates the denial of negatively perceived personality traits (Bäckström, Björklund, & Larsson, 2009; Viswesvaran & Ones, 1999). In work settings, participants can successfully simulate social desirable responses, such as positive attributes for specific job descriptions (Bagby & Marshall, 2003; Furnham, 1990; Pauls & Crost, 2004; Retzlaff, Sheehan, & Fiel, 1991; Scandell & Wlazelek, 1996). To address the vulnerability of my studies to response distortion, I will implement a scale to detect social desirability. This scale is based conceptually on detection strategies, for example, the assumption that respondents who score significantly above the norm on items about socially desirable qualities might be overstating their positive self-presentation.

Need for Chaos

Although multiple psychological motivations shape the spread of rumors and stereotypes in general (DiFonzo & Bordia, 2007), some evidence suggests that the sharing of hostile rumors and negative stereotypes about other groups specifically relates to states of conflict between the target group and the group of the rumor sharer (Laustsen & Petersen, 2015). Rumor and stereotype sharing is in part motivated by perceptions of intergroup conflict (Tooby & Cosmides, 2010). In this perspective, the person posting the hostile remarks is less concerned with the truth and more concerned with the value of the rumor to aiding in their side "winning" the intergroup conflict. Additionally, people who post hostile rumors and stereotypes may be motivated by what Petersen, Osmundsen, and Arceneaux (2018) term, "chaotic" motivations. That is, when people share hostile rumors and stereotypes they might do so with the motivation to mobilize the audience against the entire social order, rather than aiding one group within the system against another. It is possible that someone with such a need for chaos may have sympathy for "trolling" behaviors, and therefore may view "trolling" behavior as non-maliciously intentioned. Therefore, I implemented the Need for Chaos Scale (Petersen, et al., 2018) to measure need for chaos and control for this possible sympathy.

Racial Attitudes

The fundamental nature of White North American attitudes towards Black people as overtly negative is largely considered to no longer be socially acceptable. Unfortunately, negative attitudes based on race have not been eradicated, but have only grown more complex. Blatant discriminatory behaviors and prejudices are frowned upon, and people are anxious to avoid behaving in a manner that could potentially be construed as unfair or prejudiced (Fiske, 1998; Gaertner & Dovidio, 1986; Plant & Devine, 1998). One's personal prejudices or biases

however, can be expressed in far more subtle ways (Gaertner & Dovidio, 1981). Although most individuals in contemporary North American society face a strong societal and cultural demand to endorse egalitarian principles, discrimination still exists. It has been demonstrated quite convincingly in different laboratory settings, such as in the case of helping behavior in both emergency (Gaertner & Dovidio, 1977) and nonemergency situations (Gaertner & Dovidio, 1986). Although behavior that is overtly prejudiced or discriminatory is socially unacceptable and most individuals therefore consciously avoid and control explicit expression of prejudice in their responses in interracial situations, implicit and more subtle biases are still common (Devine, 1989, Greenwald, McGhee & Schwartz, 1998).

Implicit prejudice has been shown to impact the discriminatory behavior of aversive racists (Son Hing, Chung-Yan, Grunfeld, Robichaud & Zanna, 2005), and also to predict the level of bias that independent observers and Black confederates themselves perceive in the nonverbal behaviors of a White participant engaging in interracial interactions in the lab (Dovidio, Kawakami, & Gaertner, 2002). Large discrepancies between White individuals' positive explicit attitudes and negative implicit attitudes towards Black people are therefore common. In one demonstration of aversive racism (Gaertner & Dovidio, 1977), White participants witnessed a staged emergency involving either a Black or a White target, and were either under the impression that they were the only witness to the emergency, or that there were other White witnesses as well. When the participants assumed that they were the only witness in the situation, they frequently rushed to help both the Black and White targets. There was no indication of overt racism or a bias for the White target in that situation. In fact, they helped the Black target more often than the White target (94% vs. 81%, respectively). However, when they thought other witnesses were present as well, the participants were far less likely to help the

Black target than the White target (38% vs. 75% of the time). The researchers assumed that, when others were present, participants were able to rationalize their reluctance to help using factors unrelated to the target's race, and thus safely engage in discrimination against the Black target. In online contexts, where the presence of many other witnesses is assumed due to the public nature of the internet, it is quite possible that these effects will allow individuals to rationalize discrimination against targets of racial prejudice. To attempt to control for this potential bias, I will therefore employ two measures of racial attitudes.

Miller and Saucier (2018) created the Propensity to Make Attributions to Prejudice Scale (PMAPS) to assess these tendencies. The PMAPS has been shown to predict attributions to prejudice in a variety of situations, particularly ambiguous situations where behavior can be attributed to factors unrelated to racial differences (e.g., Miller et al., 2017; Stratmoen, Lawless, & Saucier, 2019). Additionally, PMAPS may be negatively associated with motivations to protect the existing social hierarchy and with anger when historically lower-status groups (i.e., Black people) claim discrimination (Miller et al., 2017). The Attitudes Toward Blacks (ATB) Scale was constructed by Brigham (1993) to measure White people's racial attitudes toward Black people in four central areas, including feelings of social distance or discomfort interacting with Black people, negative affective reactions to Black people, governmental policy (e.g., open housing, equality), and personal worry about being denied a job or promotion due to preferential treatment for Black people (based on affirmative action programs). The ATB Scale has been used recently in studies as a measure of prejudice in social mobility and economic policies (e.g., Bianchi, Hall, & Lee, 2018; Mandalaywala, Amodio, & Rhodes, 2018) and in studies that demonstrate the effectiveness of third-party confrontation on decreasing prejudice (e.g., Czopp, Monteith, & Mark, 2006).

Overview of Current Studies

Some previous research indicates that online anonymity is negatively associated with perceived honesty (e.g., Bergstrom, 2011; Reader, 2012). This is consistent with a "Trolling Hypothesis", wherein people tend to think of online anonymity as a cover for dishonest and attention-seeking aggression. More recent research has begun to examine perceptions of racism in anonymous versus identifiable online contexts (Lawless & Saucier, in preparation). Lawless and Saucier compared perceptions of overtly racist, implicitly racist, and racially neutral statements posted on apparently identifiable Facebook profiles versus on (at the time) totally anonymous Yik Yak walls. Lawless and Saucier found that, generally, the people posting the statements on the identifiable platform were rated as more racist, more honest, and as trying to convince other people at higher rates than people posting the statements on the anonymous platform. These results appear to be in favor of the Trolling Hypothesis, with anonymous people being rated as less honest and less racist.

In contrast, previous research has also shown that adolescents view anonymous cyberbullying as more severe and painful than identifiable cyberbullying (Sticca & Perren, 2012). These and other results (e.g., Bargh et al., 2002; Tidwell & Walther, 2002) could indicate that online anonymity is positively associated with perceived honesty, which would make anonymous comments more hurtful than identifiable comments. This is consistent with a "Disinhibition Hypothesis", wherein people tend to think of online anonymity as a tool for people to use for protection while they espouse what they really believe.

These competing hypotheses both have empirical support. This may be because the research supporting each hypothesis differs in terms of who is targeted by the negative online statements. In much of the research supporting the Trolling Hypothesis (e.g., Lawless & Saucier,

in preparation; Reader, 2012), the target of the negativity was a group at large (e.g., Black people in general), not an individual person, and the results of these studies indicated anonymous commentary may have been seen as less honest. In contrast, much of the research supporting the Disinhibition Hypothesis featured an individual person as the target (e.g., Bargh et al., 2002; Sticca & Perren, 2012), and the results indicated that anonymous commentary may have been seen as more honest. It is possible that the individuality of the target of the negativity at least partially explains the differing results above. The internet is rife with generalized anonymous prejudice (Hawdon, et al., 2017; Räsänen, et al., 2016; Sponholz, 2018), and it is possible that people have desensitized to it and learned to dismiss such anonymous speech as dishonest. It simply appears common for people to say extreme, antisocial things anonymously on the internet with no fear of consequences for doing so. Therefore, in an extension of previous research, I examined the effects of specificity of target on perceptions of honesty in anonymous online prejudiced speech. Further, I examined the effects of publicity of statement (i.e., a public social media post versus a private message) on these perceptions.

Chapter 2 - Study 1

Study 1 examined the effects of specificity of target of online prejudice on perceptions of the honesty, prejudice, and genuineness of intention of the prejudiced speech. Interestingly, there are two competing hypotheses in the current studies. The first hypothesis, the Trolling Hypothesis, states that anonymous statements will be rated as less honest, prejudiced, and genuine than identifiable statements because the statements are thought of as being provoked not by genuine feeling, but by the thrill of being allowed to broadcast socially unacceptable statements without the consequences that would come with being identifiable. The second hypothesis, the Disinhibition Hypothesis, states that anonymous statements will be rated as more honest, prejudiced, and genuine than identifiable statements because anonymity is thought of as merely eliminating the social pressures that usually inhibit expressions of prejudice (e.g., Spears & Lee, 1992). These hypotheses build upon previous research on anonymous online behavior that has found seemingly conflicting evidence that anonymous comments are regarded as less honest (e.g., Lawless & Saucier, in preparation; Reader, 2012) or more honest (e.g., Bargh et al., 2002; Sticca & Perren, 2012) than identifiable speech. I predicted that this conflict arises from the differences in the target of the comments. As such, I predicted that when the target is a group of people generally, anonymous comments are rated as less honest, prejudiced, and genuine than identifiable comments. However, when the target is a specific person, anonymous comments are rated as more honest, prejudiced and genuine than identifiable comments. These target effects would explain the seemingly conflicting evidence found in previous research. Specifically, I conducted a study in which participants rated the level of prejudice, honesty, and genuineness they perceived in racist comments directed toward Black people as a group versus racist comments directed at single Black individuals. I presented these comments on both identifiable

(i.e., Facebook, where fake names and profile pictures are attached to each comment) and anonymous (i.e., Reddit, where only anonymous screennames are used) social media platforms. I then examined the differences in participants' perceptions based on both anonymity of platform and specificity of target.

Study 1 Method

Participants

Participants consisted of 177 volunteers who were recruited from Amazon's Mechanical Turk software and participated in exchange for a small monetary compensation (i.e., $0.25). However, of these 177 participants, 12 were removed for finishing the survey in less than three minutes and 6 were removed for failing the attention checks and/or bot captcha; therefore, I analyzed 159 participants' responses. I conducted an a priori power analysis (gPower) with an $\alpha = .05$ and power of .95. Further, the effect size which was entered into gPower was taken from Lawless and Saucier (in preparation), which showed effect sizes of approximately .20. This analysis yielded an approximate sample size of 117 participants necessary to achieve the boundaries discussed. To ensure participant anonymity, participant names were not collected and worker identification numbers were kept separately from all other study materials. Identification information was only collected for the purposes of informed consent and exchanging appropriate compensation.

Mock Social Media Posts

I used stimuli similar to Lawless & Saucier (in preparation), which used overtly racist statements targeting groups of people and manipulated the anonymity of the person posting the statements by placing the statements in mock Facebook (identifiable) or Yik Yak (anonymous) posts. In the time since those studies, Yik Yak has changed its policy and is no longer totally

anonymous. Therefore, in the current study, I used mock Reddit posts in the anonymous conditions. Reddit is a website that allows users to post content behind any screenname they want and does not require any link to an identifiable Facebook, Google, or email account. Thus, though there is a screenname attached to posts on Reddit, it is impossible to connect a non-identifiable screenname to a real-life person, and posts are therefore anonymous.

The current study examined whether perceptions of racist comments differ based on both anonymity and individuality of target. Therefore, I specifically used 10 overtly racist mock social media posts that attacked a group of people similar to those used by Lawless and Saucier (in preparation; e.g., *Black people whine and complain about being "oppressed" yet sit at home and collect welfare. It's called hard work*) as well as 10 mock posts containing overtly racist personal attacks against individuals (e.g., *Marc doesn't deserve to be on the basketball team. He's just there because he's Black*; see Appendix C for additional examples). I also included 10 mock posts containing no racial content at all as a control (e.g., *I can't believe my professor gives straight zeros for late work. Why not just have a late penalty instead?*). Posts were evenly split amongst identifiable (Facebook) and anonymous (Reddit) social media platforms and were randomized such that all participants saw 5 posts from each condition in each anonymity condition in a within-subjects design.

Individual Differences

Propensity to Make Attributions to Prejudice. To measure beliefs about the prevalence of racial prejudice, I used the Propensity to Make Attributions to Prejudice Scale (PMAPS; Miller & Saucier, 2018). The scale includes 15 items measured on a 1 (*strongly disagree*) to 9 (*strongly agree*) Likert-type scale. It includes items such as *I consider whether people's actions are prejudiced or discriminatory*. I calculated a composite score for PMAPS by reverse-scoring

antithetical items and calculating an average score for each participant with higher scores indicating greater tendencies to attribute causes of behavior to racial prejudice.

Explicit prejudice toward Black People. To measure participants' levels of explicit racial prejudice toward Black individuals, I used the Attitudes Toward Blacks (ATB; Brigham, 1993) scale. The scale includes 20 items measured on a 1 (*strongly disagree*) to 9 (*strongly agree*) scale. It includes items such as *I would rather not have Blacks live in the same apartment building I live in.* I calculated a composite score for ATB by reverse-scoring antithetical items and calculating an average score for each participant with higher scores indicating greater levels of blatant anti-Black prejudice.

Need for Chaos. To measure participants' levels of desire to fight against established social order, I used the Need for Chaos Scale (Petersen et al., 2018). The scale includes eight items measured on 1 (*strongly disagree*) to 9 (*strongly agree*) scale. It includes items such as *I think society should be burned to the ground.* I calculated a composite score for Need for Chaos by calculating an average score for each participant with higher scores indicating greater levels of desire for chaos.

Social Desirability. To measure participants' tendencies toward socially desirable behavior, I used the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), which defines social desirability as the need for social approval. This instrument includes 33 items, which are to be classified as true or false by the respondent. Some of these items correspond to sentences that describe desirable but uncommon daily behaviors (attribution items, scored if answered "true"; e.g., *I am always courteous, even to people who are disagreeable.*), whereas others describe highly common but socially undesirable behaviors (denial items, scored when answered "false"; e.g., *There have been occasions when I felt like smashing things*).

Therefore, social desirability was scored from 0 – 33 as the number of socially desirable responses made by the participant.

Criterion Variables

Each of the following measures was chosen to represent a specific facet of perceptions of online prejudice and the people posting it that has been discussed in previous literature. Specifically, I included items assessing the extent to which participants perceived the posts as racist and honest as well as items assessing the extent to which participants perceived the posters as racist, honest, genuine in their belief of what they have posted, attempting to convince others of what they have posted, and attempting to seek attention for attention's own sake. Each of these measures is described below, and the materials are included in Appendix D.

Perceived racial prejudice of the post. To examine the extent to which participants perceived each post as racist, I used a perceived racial prejudice item employed by Lawless and Saucier (in preparation). This item examines the extent to which participants perceive the social media post as racist. This item is measured on a 1 (*not at all*) to 9 (*very much*) scale, with higher ratings indicating greater levels of perceived racial prejudice of the post.

Perceived maliciousness of the post. To examine the extent to which participants perceived each post as malicious, I used three perceived maliciousness items (e.g., *This post is meant to harm*). These items examine the extent to which participants perceive the social media post as malicious. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. I calculated a composite score for the perceived maliciousness of the post by reverse scoring antithetical items and calculating an average score with higher scores indicating greater levels of perceived maliciousness of the post.

Perceived honesty of the post. To examine the extent to which participants perceived each post as honest, I used two perceived honesty items similar to those used by Lawless and Saucier (in preparation). These items examine the extent to which participants perceive the social media post as honest. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. I calculated a composite score for the perceived honesty of the post by calculating an average score with higher scores indicating greater levels of perceived honesty of the post.

Perceived racial prejudice of the person posting. To examine the extent to which participants perceived the person posting each statement as racist, I used a perceived racial prejudice item employed by Lawless and Saucier (in preparation). This item examines the extent to which participants perceive the person posting the social media post as racist. This item is measured on a 1 (*not at all*) to 9 (*very much*) scale, with higher ratings indicating greater levels of perceived racial prejudice of the person posting the statement.

Perceived maliciousness of the person posting. To examine the extent to which participants perceived each poster as malicious, I used three perceived maliciousness items (e.g., *The person who posted this intended to harm the person(people) this post is about.*). These items examine the extent to which participants perceive the person posting the social media post as malicious. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. I calculated a composite score for the perceived maliciousness of the poster by reverse scoring antithetical items and calculating an average score with higher scores indicating greater levels of perceived maliciousness of the poster.

Perceived honesty of the person posting. To examine the extent to which participants perceived each poster as honest, I used two perceived honesty items similar to those used by Lawless and Saucier (in preparation). These items examine the extent to which participants

perceive the poster as honest. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. I calculated a composite score for the perceived honesty of the poster by calculating an average score with higher scores indicating greater levels of perceived honesty of the poster.

Perceived attention seeking of the person posting. To examine the extent to which participants perceived each poster as seeking attention *rather than* being honest, I used a perceived attention-seeking item similar to that employed by Lawless and Saucier (in preparation; *i.e, The person who posted this is simply looking for attention*). This item examines the extent to which participants perceive the poster as seeking attention. This item is measured on a 1 (*not at all*) to 9 (*very much*) scale, with higher ratings indicating greater levels of perceived attention-seeking of the poster.

Procedure

The current study was conducted online using Amazon's Mechanical Turk software. Once participants signed up, they followed a link to my study on Qualtrics. Participants gave informed consent prior to participation. After providing demographic information (e.g., sex, race, age), participants read and responded to all 30 mock social media posts in the randomized fashion described above. Participants were debriefed after they completed the study to allow the experimenters to answer any questions the participants had.

Results

Following the cleaning of my dataset (e.g., removing participants who completed the questionnaire in less than three minutes, removing participants who failed the bot captcha), I computed composite scores for each of my continuous variables. As noted in the Materials section, for each of the measures, I averaged participants' scores on each individual item after

reverse scoring antithetical items to create composite scores. On each measure, higher scores represent higher levels of the construct being measured.

I examined the bivariate correlations among my predictor variables (see Table 1). Consistent with previous research (e.g., Miller & Saucier, 2018) and my hypotheses, PMAPS and ATB were negatively correlated (r = -.65), and PMAPS and SD were not significantly correlated (r = .11). Additionally, social desirability was negatively correlated with ATB (r = -.58) and Need for Chaos (r = - .36). However, these correlations are not central to the main hypotheses of the current studies, so I will not be discussing them further.

I then examined the bivariate correlations among the criterion variables: perceived racial prejudice of the post, maliciousness of the post, honesty of the post, racial prejudice of the person posting, maliciousness of the person posting, honesty of the person posting, and attention seeking of the person posting (see Table 2). Consistent with previous research by Lawless and Saucier (in preparation), there were positive correlations between the perceived racial prejudice of the post, maliciousness of the post, racial prejudice of the person posting, and maliciousness of the person posting (see Table 2). There was also a positive correlation between perceived honesty of the post and honesty of the person (r = .46). In addition, there was a negative correlation between perceived honesty of the person posting and perceived attention seeking of the person posting (r = -.27).

I then tested whether anonymous posts are seen as less racist, less honest, and more attention seeking than similar identifiable posts. Recall, there were two competing hypotheses founded on previous research. The first hypothesis, the Trolling Hypothesis, states that anonymous statements will be rated as less honest and prejudiced, and more attention-seeking than identifiable statements because the statements are thought of as being provoked not by

genuine feeling, but by the thrill of being allowed to broadcast socially unacceptable statements without the consequences that would come with being identifiable. The second hypothesis, the Disinhibition Hypothesis, states that anonymous statements will be rated as more honest and prejudiced, and less attention-seeking than identifiable statements because anonymity is thought of as removing the social pressures that usually inhibit genuine expressions of prejudice. Additionally, I predicted that when the target is a group of people generally, anonymous comments would be rated as less honest and prejudiced, and more attention-seeking than identifiable comments because it is seen as the person posting to disparage entire groups, not as an act of genuine hatred, but because it is thrilling to participate in the taboo act of espousing prejudiced rhetoric. However, when the target is a specific person, I hypothesized anonymous comments would be rated as more honest and prejudiced than identifiable comments because the person posting has set out to disparage an individual by name and therefore may appear to have a personal vendetta fueled by genuine feeling toward the target individual. These target effects would explain the seemingly conflicting evidence found in previous research (e.g., Bargh et al., 2002; Sticca & Perren, 2012; Reader, 2012).

To test these hypotheses against one another, I conducted a series of multilevel model analyses predicting the criterion variables and including PMAPS, Need for Chaos, Racial Content of the Post, Anonymity, Singularity of Target, and the interaction between Anonymity and Singularity of Target as predictor fixed effects (see Tables 3-9), and allowing participants' intercepts to vary. Consistent with my hypotheses, there were significant unique effects of PMAPS (*Prejudice of Post: F*(1, 158) = 97.21, $p < .001$; *Maliciousness of Post*: *F*(1, 158) = 123.70, $p < .001$, *Prejudice of Person: F*(1, 158) = 145.61, $p < .001$, *Maliciousness of Person F*(1, 158) = 134.21, $p < .001$; see PMAPS β values in Tables 3-6) such that, generally people

higher in PMAPS viewed posts and people as both more prejudiced and more malicious. There were also significant unique effects of Racial Content (*Prejudice of Post: F*(1, 158) = 123.46, $p < .001$; *Maliciousness of Post*: *F*(1, 158) = 98.52, $p < .001$, *Prejudice of Person: F*(1, 158) = 98.53, $p < .001$, *Maliciousness of Person F*(1, 158) = 76.52, $p < .001$; see Racial Content β values in Tables 3-6) such that posts containing racial content were rated as generally more malicious and more prejudiced. Also, there were significant unique effects of Need for Chaos (*Maliciousness of Post*: *F*(1, 158) = 18.07, $p < .001$, *Maliciousness of Person F*(1, 158) = 12.61, $p < .001$; see Need for Chaos β values in Tables 3-6) such that, generally people higher in Need for Chaos viewed posts and people as less malicious, perhaps because need for chaos is associated with wanting to buck the social order, potentially leading to sympathizing with online behavior that does so.

Additionally, consistent with the Trolling hypothesis, there were significant unique effects of Anonymity (*Prejudice of Person: F*(1, 786) = 8.24, $p = .004$, *Maliciousness of Person F*(1, 786) = 5.64, $p = .018$, *Honesty of Post: F*(1, 786) = 9.36, $p = .001$, *Honesty of Person: F*(1, 786) = 8.57, $p = .002$; see Anonymity β values in Tables 3-9), such that people posting anonymous posts were rated as less prejudiced, malicious, and honest than people posting identifiably. This could suggest that people view anonymous posts as trolling, not meant to be taken seriously or as truth, but rather intended to garner extreme reactions by bucking against the social order. Additionally, there were significant effects of Singularity of Target (*Maliciousness of Post*: *F*(1, 786) = 4.36, $p = .042$, *Maliciousness of Person F*(1, 786) = 8.74, $p = .039$, *Honesty of Post: F*(1, 786) = 10.45, $p < .001$, *Honesty of Person: F*(1, 786) = 11.86, $p < .001$; see Singularity β values in Tables 3-9), such that posts targeting singular individuals were rated as more malicious and honest than posts targeting Black people as a whole. This suggests that posts targeting individuals are not seen as trolling, perhaps because of the personal connection

suggested by targeting a named individual. These main effects were qualified by significant two-way interactions between Anonymity and Singularity of Target (*Honesty of Post: F*(1, 786) = 7.54, *p* = .006, *Honesty of Person: F*(1, 786) = 5.97, *p* = .013, *Attention Seeking*: *F*(1, 786) = 6.34, *p* = .009; see interaction term β values in Tables 3-9). These interactions indicate that the effects of anonymity of post on these criterion variables depended upon whether the target was a group of people or a named individual.

I then conducted simple slopes analyses on the interaction terms that were significant to determine whether my final hypotheses were supported (see Table 10). As predicted, anonymous posts that targeted singular named individuals were rated as more honest and attention-seeking than identifiable posts targeting either named individuals or a group as a whole. However, anonymous posts targeting groups were rated as less honest, and more attention-seeking than other types of posts. This could be because people who post anonymously espousing prejudiced rhetoric against large groups of people are colloquially known online as trolls and are thought of as posting purely for the thrill of espousing taboo prejudice rather than out of genuine belief.

Chapter 3 - Study 2

Method

Participants

Participants consisted of 169 volunteers who were recruited from Amazon's Mechanical Turk software and participated in exchange for a small monetary compensation (i.e., $0.25). However, of these 169 participants, 9 were removed for finishing the survey in less than three minutes and 7 were removed for failing the attention checks and/or bot captcha; therefore, I analyzed 153 participants' responses. I conducted an a priori power analysis (gPower) with an $\alpha = .05$ and power of .95. Further, the effect size which was entered into gPower was taken from Lawless and Saucier (in preparation), which showed effect sizes of approximately .20. This analysis yielded an approximate sample size of 122 participants necessary to achieve the boundaries discussed. To ensure participant anonymity, participant names were not collected and worker identification numbers were kept separately from all other study materials. Identification information was only collected for the purposes of informed consent and exchanging appropriate compensation.

Mock Social Media Posts and Messages

I used the content from 8 the same 10 overtly racist mock social media posts that were used in Study 1. Content was presented in a 2 (identifiable/anonymous sender) x 2 (public post/private message) x 2 (attacking a particular person/attacking a group) within-subjects design. That is, posts were evenly split amongst identifiable (Facebook) and anonymous (Reddit) social media platforms and were randomized such that all participants saw posts from both target conditions in each anonymity condition and privacy condition. Posts and messages appeared to be on either an identifiable (Facebook) or anonymous (Reddit) social media platform, and were

either public (on a Facebook wall or personal Subreditt) or private (in a private Facebook Message or Reddit Private Message). All manipulations were presented to participants in a randomized fashion.

Individual Differences and Criterion Variables

Individual differences and criterion variables were the same as those in Study 1. Individual differences included: PMAPS, ATB, Social Desirability, and Need for Chaos. Criterion variables included: perceived racial prejudice of the post, perceived maliciousness of the post, perceived honesty of the post, perceived racial prejudice of the person posting, perceived maliciousness of the person posting, perceived honesty of the person posting, and perceived attention seeking of the person posting. Each of these measures was chosen to represent a specific facet of perceptions of online racial dialogue and the people posting it that has been discussed in previous literature. Materials are included in Appendix D.

Procedure

The current study was conducted online using Amazon's Mechanical Turk software. Once participants signed up, they followed a link to my study on Qualtrics. Participants gave informed consent prior to participation. After providing demographic information (e.g., sex, race, age), participants read and responded to all 16 mock social media posts in the randomized fashion described above. Participants were debriefed after they completed the study to allow the experimenters to answer any questions the participants had.

Results

Following the cleaning of my dataset (e.g., removing participants who completed the questionnaire in less than three minutes, removing participants who failed the bot captcha), I computed composite scores for each of my continuous variables. As noted in the Materials

section, for each of the measures, I averaged participants' scores on each individual item after reverse scoring antithetical items to create composite scores. On each measure, higher scores represent higher levels of the construct being measured.

I examined the bivariate correlations among my predictor variables (see Table 11). Consistent with previous research (e.g., Miller & Saucier, 2018) and my hypotheses, PMAPS and ATB were negatively correlated (r = -.68) and PMAPS and SD were not significantly correlated (r = .13). Additionally, social desirability was negatively correlated with ATB (r = -.46) and Need for Chaos (r = -. 43). However, these correlations are not central to the main hypotheses of the current studies, so I will not be discussing them further.

I then examined the bivariate correlations among the criterion variables: perceived racial prejudice of the post, maliciousness of the post, honesty of the post, racial prejudice of the person posting, maliciousness of the person posting, honesty of the person posting, and attention seeking of the person posting (see Table 12). Consistent with previous research by Lawless and Saucier (in preparation), there were positive correlations between the perceived racial prejudice of the post, maliciousness of the post, racial prejudice of the person posting, and maliciousness of the person posting (see Table 12). There was also a positive correlation between perceived honesty of the post and honesty of the person (r = .68). In addition, there was a negative correlation between perceived honesty of the person posting and perceived attention seeking of the person posting (r = -.31).

I then tested whether anonymous posts are seen as less racist, less honest, and more attention seeking than similar identifiable posts. I predicted results similar to those in Study 1. Again, I predicted that when the target is a group of people generally, anonymous comments will be rated as less honest and prejudiced, and more attention-seeking than identifiable comments

because it is seen as the person posting to disparage entire groups, not as an act of genuine hatred, but because it is thrilling to participate in the taboo act of espousing prejudice. However, when the target is a specific person, I hypothesized anonymous comments would be rated as more honest and prejudiced than identifiable comments because the person posting has set out to disparage an individual by name and therefore may appear to have a personal vendetta fueled by genuine feeling toward the target individual. However, I expected that the effects of anonymity would be mitigated by the privacy of the message. That is, when the post is public, I would find the above effects; however, when the post is private, it would be rated as equally honest and prejudiced regardless of anonymity. Trolling is predicated on attention-seeking, and private messages do not typically garner the amount of attention a troll is looking for. Therefore, it is possible that anonymous private messages will not be thought of as trolling and will therefore be perceived as equally honest as identifiable messages.

To test these hypotheses against one another, I conducted a series of multilevel model analyses predicting the criterion variables and including PMAPS, Need for Chaos, Anonymity, Singularity of Target, Privacy of Message, and the interactions between Anonymity, Singularity of Target, and Privacy of Message as predictor fixed effects, and allowing participants' intercepts to vary.

Consistent with Study 1 and my hypotheses, there were significant unique effects of PMAPS (*Prejudice of Post:* $F(1, 152) = 96.32$, $p < .001$; *Maliciousness of Post*: $F(1, 152) = 102.74$, $p < .001$, *Prejudice of Person:* $F(1, 152) = 98.52$, $p < .001$, *Maliciousness of Person* $F(1, 152) = 89.58$, $p < .001$; see PMAPS β values in Tables 13-16) such that, generally people higher in PMAPS viewed posts and people as both more prejudiced and more malicious. Again, there were also significant unique effects of Need for Chaos (*Maliciousness of Post*: $F(1, 152) = 19.23$,

$p < .001$, *Maliciousness of Person* $F(1, 152) = 7.18$, $p = .049$; see Need for Chaos β values in Tables 3-6) such that, generally people higher in Need for Chaos viewed posts and as less malicious, perhaps because Need for Chaos is associated with sympathizing with people who espouse rhetoric that goes against established social rules.

Additionally, again consistent with the Trolling hypothesis and with Study 1, there were significant unique effects of Anonymity (*Prejudice of Person:* $F(1, 302) = 8.24$, $p < .001$, *Maliciousness of Person* $F(1, 302) = 5.64$, $p = .018$, *Honesty of Post:* $F(1, 302) = 8.43$, $p = .002$, *Honesty of Person:* $F(1, 302) = 10.46$, $p < .001$; see Anonymity β values in Tables 13-19), such that people posting anonymous posts were rated as less prejudiced, malicious, and honest than people posting identifiably. This could suggest that people view anonymous posts as trolling, not meant to be taken seriously or as truth, but rather intended to garner extreme reactions by bucking against the social order. Additionally, there were significant effects of Singularity of Target (*Maliciousness of Post*: $F(1, 302) = 4.76$, $p = .048$, *Maliciousness of Person* $F(1, 302) = 8.74$, $p < .001$, *Honesty of Post:* $F(1, 302) = 9.95$, $p < .001$, *Honesty of Person:* $F(1, 302) = 80.95$, $p = .042$; see Singularity β values in Tables 13-19), such that posts targeting singular individuals were rated as more malicious and honest than posts targeting Black people as a whole. There were also significant effects of Privacy of Message (*Honesty of Post:* $F(1, 302) = 10.84$, $p < .001$, *Honesty of Person:* $F(1, 302) = 11.25$, $p < .001$; see Privacy β values in Tables 13-19), such that posts sent as private messages were rated as more malicious and honest than posts made publicly. Taken together, these results suggest that posts targeting individuals or sent as private messages are not seen as trolling, perhaps because of the personal connection suggested by targeting a named individual or sending a personal message. These main effects were qualified by significant two-way interactions between Anonymity and Singularity of Target

(*Honesty of Post: F*(1, 302) = 9.78, *p* < .001, *Honesty of Person: F*(1, 302) = 6.01, *p* = .009, *Attention Seeking*: *F*(1, 302) = 5.87, *p* = .013; see interaction term β values in Tables 13-19) as well as between Anonymity and Privacy of Message (*Honesty of Post: F*(1, 302) = 5.51, *p* = .017, *Honesty of Person: F*(1, 302) = 6.22, *p* = .007, *Attention Seeking*: *F*(1, 302) = 3.98, *p* = .039; see β values in Tables 13-19), These interactions indicate that the effects of anonymity of post on these criterion variables depended upon whether the target was a group of people or a named individual and whether the post was made publicly or in a private message.

I then conducted simple slopes analyses on the interaction terms that were significant to determine whether my final hypotheses were supported (see Tables 20 and 21). As predicted, and consistent with Study 1, anonymous posts that targeted singular named individuals were rated as more honest than identifiable posts targeting either named individuals or a group as a whole. However, anonymous posts targeting groups were rated as less honest, and more attention-seeking than other types of posts. Additionally, posts made as private messages were rated as similarly honest and attention-seeking regardless of anonymity, but public posts were rated as more attention seeking and less honest when they were anonymous. Taken together, these results suggest that only public posts targeting groups of people are considered trolling. Adding a personal connection, either by targeting a named individual or by making a post via private message, negates the effects anonymity seems to have on perceptions of public, non-personal posts.

Chapter 4 - General Discussion

In these two studies, I have begun to clarify the circumstances under which anonymous online behavior is viewed as honest and/or attention seeking, which contributes to the existing literature on online behavior, and elucidates some of the potential differences between online and traditional face-to-face interaction. Some previous research has suggested online anonymity is negatively associated with perceived honesty (e.g., Lawless & Saucier, in preparation; Reader, 2012). This is consistent with my Trolling Hypothesis, wherein people tend to think of online anonymity as a cover for baseless aggression. In contrast, some previous research has suggested online anonymity is positively associated with perceived honesty (e.g., Sticca & Perren, 2012), which is consistent with my Disinhibition Hypothesis, wherein people tend to think of online anonymity as a tool people use to protect themselves from potential social consequences of espousing their own genuine beliefs. I contend that it is possible that each of these seemingly competing hypotheses can explain differing perceptions based on who is being targeted by negative statements online and who is the intended audience. In literature that supports the Trolling Hypothesis, the target of the negative behavior is typically a group of people as a whole (e.g., Black people in general), and the intended audience is also a large group of people (e.g., an entire online forum, or a public Facebook audience; e.g., Lawless & Saucier, in preparation; Reader, 2012). In contrast, in literature that supports the Disinhibition Hypothesis, the target of the behavior is typically an individual, non-celebrity person and the audience is typically much smaller (e.g., cases of cyberbullying a particular classmate; e.g., Dilmac, 2009; Price & Dalgleish, 2010; Sticca & Perren, 2012). It is possible that this individuation of the target and size variation of intended audience at least partially explain the differing results in previous literature.

Consistent with my hypotheses, anonymous posts that targeted singular named individuals were rated as more honest than identifiable posts. This was consistent with my Disinhibition Hypothesis, and with past literature with similar findings (e.g., Dilmac, 2009; Price & Dalgleish, 2010; Sticca & Perren, 2012). This could be because anonymity is thought of as removing the social pressures that usually inhibit genuine expressions of racism. When the target is a specific person, anonymous comments were rated as more honest than identifiable comments, perhaps because the person posting has set out to disparage an individual by name and therefore may appear to have a personal vendetta toward the target individual. Because the online environment involves anonymity, invisibility, and lack of face-to-face contact, punishment and repercussions are considered less likely to occur as compared with the offline world (Suler, 2004), and it is possible that people believe others take advantage of what might be called the "Wild Wild Web" to spread genuine hatred. Additionally, deindividuation increases in situations with greater anonymity and leads to a breakdown of traditional social contracts (Postmes et al., 2001). A lack of distinguishing features online alters perceptions of the self and of others as individuals (Postmes & Baym, 2005). This deindividuation can lift the social ban on racism, and it is possible that third-party observers understand how anonymity can allow hatred to manifest in an online context, leading them to perceive interpersonal expressions of prejudice as more honest when the perpetrator is anonymous.

Interestingly, people posting anonymously were rated as similarly prejudiced to identifiable people in the current studies, which is inconsistent with some past literature (e.g., Lawless & Saucier, in preparation). This discrepancy with past literature could be due to the addition of the singularity of target and private message conditions. It is possible that, because many of the posts were clearly directed at or meant for a particular individual, even anonymous

writers were rated as prejudiced. Perhaps participants perceived the act of targeting an individual, either via a named post or via a private message, as inherently prejudiced, whether the perpetrator was anonymous or identifiable. This would be consistent with past literature on phenomena such as cyberbullying where the target is clearly identified (e.g., Bargh et al., 2002; Sticca & Perren, 2012).

Consistent with my hypotheses, anonymous posts targeting groups were rated as less honest, and more attention-seeking than all other posts. This was consistent with my Trolling Hypothesis, and with past literature (e.g., Lawless & Saucier, in preparation; Reader, 2012). Trolls are thought of as posting purely for the thrill of breaking taboos rather than out of genuine belief. Additionally, the internet is rife with generalized anonymous prejudice (e.g., Li, 2005; Willard, 2007), and it is possible that people have desensitized to it and learned to dismiss such anonymous speech as dishonest. It simply appears common for people to say extreme, antisocial things anonymously on the internet with no fear of consequences for doing so. Additionally, many findings have indicated a link between aggression and attention-seeking, and engaging in cyber-bullying behaviors (Harman et al., 2005; Li, 2005; Willard, 2007). It is possible that observers understand the possibility of attention-seeking motivations, and therefore attribute online expressions of prejudice to honesty in some circumstances and to attention-seeking or trolling in others. Essentially, anonymous statements targeting groups were rated as less honest and more attention-seeking than identifiable statements because the anonymous statements are thought of as being provoked not by genuine feeling, but by the thrill of being allowed to broadcast typically socially unacceptable statements.

Additionally, consistent with my hypotheses, anonymous posts meant for a large audience (i.e., a public Facebook or Reddit post) were perceived as less honest than private

messages, perhaps because trolling is seen as an attention-seeking activity. If someone sends prejudiced insults to a single person privately, they may not be seeking widespread attention like a troll would, and may therefore be seen as genuine. Trolls frequently perform anti-social and widely unacceptable behaviors, such as espousing racist or sexist hate speech. Trolling is typically predicated on sensationalism and emotional exploitation, both of which can be met via extreme expressions of prejudice, particularly against large groups of people and in view of a large audience. Therefore, it is possible that people perceive anonymous expressions of prejudice targeting groups of people as attention-seeking behavior rather than as honest dissemination of personal beliefs, particularly when such speech is posted publicly. However, it is important to note that the perception of these posts as dishonest does not necessarily mean they are harmless. The sleeper effect is a phenomenon wherein, whereas people are typically not persuaded immediately by a non-credible source, they become more persuaded after the passage of time, likely because they have forgotten the source and how non-credible it was (e.g., Kumkale & Abarracín, 2004). This effect may mean that even anonymous posts online can persuade people over time, making them more insidious than they may immediately appear.

Limitations and Future Directions

The current studies are not without limitations. The first limitation is the cross-sectional nature of the current studies. This limits my ability to draw causal conclusions about the relationships between current political climate, participants' levels of racial prejudice, and their perceptions of online racial prejudice. One could make the argument that participants' levels of racial prejudice are relatively consistent across time. However, I would be hesitant to draw concrete causal conclusions from the proposed studies, particularly given the recent uptick in publicized prejudiced speech in the United States' current political climate (Crandall, Miller, &

White, 2018; Waltman, 2018). It is possible that this recent lift of prejudice suppression might lead participants to believe that online prejudice is more honest than they would in a different climate because public racial prejudice has become more salient in the news over the past few years. Steps should be taken in the future to extend this research by examining perceptions of online behavior under differing political and social climates.

An additional limitation in the current studies is the usage of mock posts that are free of other common factors (e.g., rebutting comments, likes, upvotes), potentially harming the ability of my studies to generalize. In conducting studies in this fashion, participants are not able to perceive online social cues and other indications surrounding the intent of the perpetrator of the racism, or the interpretation of the speech by the target and online community. Instead, they are given an ambiguous post by itself and asked to report their perceptions. That said, there are ethical concerns in the employment of more realistic procedures (e.g., using posts actually found online). As such, there are limitations to the generalizability of the current studies to real situations. Third party observers may react differently if they were to see the comments and reactions of the perpetrator, target, and online community. Thus, the results of the current studies may not generalize to real world events. Future studies should add more realistic online interactions and could manipulate community reactions to prejudice by manipulating the number of likes or upvotes a post garners or by adding confirming or disavowing comments to the posts.

Additionally, it is possible that participants from an online environment (i.e., MTurk) have varying levels of experience with expressions of prejudice online. I did not ask participants what online communities they frequent or how much time they spend online. It is possible that someone who only frequents communities that discourage and ban prejudice (e.g., Nerdfighteria, r/Wholesome) would perceive online expressions of prejudice differently from someone who

frequents communities that allow expressions of prejudice in the form of humor or in invitations of debate (e.g., r/TheDonald). In future studies, I would like to examine and control for the types of internet use participants frequently engage in in order to explore the possible effects online community exposure can have on perceptions of online prejudice.

Conclusion

Across two studies, I examined the effects of anonymity, singularity of target, and privacy of message on third-party perceptions of the honesty of online prejudiced speech. These studies are timely and extend the existing literature on internet behavior by further examining the relationships between various possible online social conditions and community reactions based on those conditions. The potential implications of the current studies may be that factors of the online environment, such as anonymity of platform, affect how individuals react to online prejudiced speech. Specifically, people may disregard anonymous expressions of prejudice that are made against groups as a whole as trolling, dismissing them as dishonest and potentially not worth "feeding" (i.e., fighting against). These studies demonstrate that many people may not take online prejudiced rhetoric seriously, particularly when it is made anonymously, which could foster toxic online environments that are conducive to cyberbullying and even incitements to real-world violence against marginalized groups. This may especially be true in communities that thrive off anonymity (e.g., Reddit, Whispr) or allow anonymous usernames (e.g., YouTube, online gaming platforms). Therefore, these and future studies along this line of research are important to fully understand the factors at play within internet culture. Whereas the internet has a unique ability to bring people together in truly global communities, it also may have the potential to foster putrid communities based on deindividuated hatred.

Chapter 5 - Tables

Table 1

Means, Standard Deviations, and Bivariate Correlations between Predictor Variables in Study 1

Variable	*M (SD)*	1	2	3	4
1. Social Desirability	13.72 (5.78)	(.87)			
2. ATB	3.22 (1.32)	-.58***	(.90)		
3. PMAPS	5.88 (1.53)	.11	-.65***	(.88)	
4. Need for Chaos	4.04 (2.27)	-.36**	.30**	-.27**	(.85)

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 2

Means, Standard Deviations, and Bivariate Correlations between Criterion Variables in Study 1

Variable	*M (SD)*	1	2	3	4	5	6	7
1. Post Prejudiced	6.54 (1.34)	-						
2. Post Malicious	5.60 (1.32)	.58***	-					
3. Post Honesty	4.97 (1.53)	.11	-.13	-				
4. Person Prejudiced	5.87 (1.26)	.75***	.34**	.10	-			
5. Person Malicious	4.59 (1.87)	.42**	.58***	-.06	.33**	-		
6. Person Honesty	4.27 (2.06)	.21*	-.09	.46***	.19	-.03	-	
7. Attention Seeking	4.04 (2.27)	.08	.30**	-.31**	.16	.25*	-.27**	-

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 3

Fixed Effects Summary Table for Perceptions of the Post as Prejudiced

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.71	0.07	97.21	<.001
Need for Chaos	-0.06	0.05	1.56	.210
Racial Content	0.84	0.04	123.46	<.001
Anonymity	-0.11	0.04	8.49	.004
Singularity of Target	-0.02	0.04	0.21	.648
Anonymity*Singularity of Target	-0.04	0.07	0.37	.544

Table 4

Fixed Effects Summary Table for Perceptions of the Post as Malicious

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.82	0.07	123.70	<.001
Need for Chaos	-0.19	0.05	18.07	<.001
Racial Content	0.73	0.04	98.52	<.001
Anonymity	0.01	0.04	0.06	.802
Singularity of Target	0.14	0.04	4.36	.042
Anonymity*Singularity of Target	-0.17	0.08	5.19	.023

Table 5

Fixed Effects Summary Table for Perceptions of the Post as Honest

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.51	0.08	38.82	<.001
Need for Chaos	-0.05	0.04	1.18	.278
Racial Content	0.32	0.04	27.98	<.001
Anonymity	-0.12	0.04	9.36	.001
Singularity of Target	0.18	0.04	10.45	<.001
Anonymity*Singularity of Target	-0.21	0.07	7.54	.006

Table 6

Fixed Effects Summary Table for Perceptions of the Person as Prejudiced

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.78	0.06	145.61	<.001
Need for Chaos	-0.11	0.04	6.73	.010
Racial Content	0.73	0.04	98.53	<.001
Anonymity	0.11	0.04	8.24	.004
Singularity of Target	-0.01	0.04	0.12	.724
Anonymity*Singularity of Target	-0.13	0.08	3.06	.080

Table 7

Fixed Effects Summary Table for Perceptions of the Person as Malicious

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.79	0.07	134.21	<.001
Need for Chaos	-0.18	0.05	12.61	<.001
Racial Content	0.63	0.04	76.52	<.001
Anonymity	-0.09	0.04	5.64	.018
Singularity of Target	0.16	0.04	8.74	.039
Anonymity*Singularity of Target	-0.15	0.08	3.55	.060

Table 8

Fixed Effects Summary Table for Perceptions of the Person as Honest

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.63	0.08	69.06	<.001
Need for Chaos	-0.003	0.05	0.01	.941
Racial Content	0.45	0.04	37.92	<.001
Anonymity	-0.12	0.04	8.57	.002
Singularity of Target	0.21	0.04	11.86	<.001
Anonymity*Singularity of Target	-0.18	0.07	5.97	.013

Table 9

Fixed Effects Summary Table for Perceptions of Attention Seeking

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.64	0.06	109.92	<.001
Need for Chaos	0.17	0.04	19.70	<.001
Racial Content	0.58	0.04	48.52	<.001
Anonymity	0.03	0.04	0.58	.447
Singularity of Target	0.03	0.04	0.64	.425
Anonymity*Singularity of Target	-0.19	0.07	6.34	.009

Table 10

Results from Simple Slopes Analyses of Significant Interactions Between Anonymity and Singularity of Target

	Anonymity		*Singularity of Target*	
Perception	*r*	*t*	*r*	*t*
Honesty of Post	-0.37	-2.82*	0.41	3.21*
Honesty of Person	-0.35	-2.65*	0.38	2.98*
Attention Seeking	0.28	2.07*	-0.29	-2.08*

*$p \leq .05$

Table 11

Means, Standard Deviations, and Bivariate Correlations between Predictor Variables in Study 2

Variable	*M (SD)*	1	2	3	4
1. Social Desirability	14.21 (5.36)	(.82)			
2. ATB	2.96 (1.29)	-.46**	(.85)		
3. PMAPS	5.64 (1.39)	.13	-.68***	(.91)	
4. Need for Chaos	3.97 (2.18)	-.43**	.34**	-.31*	(.86)

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 12

Means, Standard Deviations, and Bivariate Correlations between Criterion Variables in Study 2

Variable	*M (SD)*	1	2	3	4	5	6	7
1. Post Prejudiced	6.54 (1.34)	-						
2. Post Malicious	5.60 (1.32)	.58***	-					
3. Post Honesty	4.97 (1.53)	.11	-.12	-				
4. Person Prejudiced	5.87 (1.26)	.68***	.34**	.07	-			
5. Person Malicious	4.59 (1.87)	.39**	.58***	-.06	.38**	-		
6. Person Honesty	4.27 (2.06)	.20*	-.09	.68***	.19	-.03	-	
7. Attention Seeking	4.04 (2.27)	.06	.30**	-.29**	.14	.25*	-.31**	-

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 13

Fixed Effects Summary Table for Perceptions of the Post as Prejudiced Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.74	0.10	96.32	<.001
Need for Chaos	-0.20	0.08	18.79	<.001
Anonymity	0.07	0.03	0.21	.653
Singularity of Target	-0.02	0.04	0.06	.819
Privacy of Message	0.03	0.05	0.12	.761
Anonymity*Singularity of Target	-0.04	0.06	0.48	.452
Anonymity*Privacy of Message	0.06	0.10	0.03	.934

Table 14

Fixed Effects Summary Table for Perceptions of the Post as Malicious Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.78	0.10	102.74	<.001
Need for Chaos	-0.21	0.08	19.23	<.001
Anonymity	0.01	0.03	0.04	.853
Singularity of Target	0.15	0.04	4.76	.048
Privacy of Message	0.12	0.05	3.84	.066
Anonymity*Singularity of Target	-0.11	0.06	3.19	.087
Anonymity*Privacy of Message	0.10	0.09	2.86	.135

Table 15

Fixed Effects Summary Table for Perceptions of the Post as Honest Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.45	0.10	33.92	<.001
Need for Chaos	0.18	0.08	12.43	<.001
Anonymity	-0.12	0.03	8.43	.002
Singularity of Target	0.16	0.04	9.95	<.001
Privacy of Message	0.19	0.05	10.84	<.001
Anonymity*Singularity of Target	-0.19	0.07	9.78	<.001
Anonymity*Privacy of Message	-0.14	0.10	5.51	.017

Table 16

Fixed Effects Summary Table for Perceptions of the Person as Prejudiced Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.75	0.10	98.52	<.001
Need for Chaos	-0.11	0.08	6.51	.018
Anonymity	-0.11	0.03	8.24	.004
Singularity of Target	-0.01	0.04	0.12	.724
Privacy of Message	0.02	0.05	0.15	.693
Anonymity*Singularity of Target	0.11	0.06	3.19	.067
Anonymity* Privacy of Message	-0.08	0.10	0.76	.253

Table 17

Fixed Effects Summary Table for Perceptions of the Person as Malicious Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.71	0.10	89.58	<.001
Need for Chaos	-0.15	0.08	7.18	.049
Anonymity	-0.14	0.03	5.64	.018
Singularity of Target	0.15	0.04	8.74	.042
Privacy of Message	0.09	0.04	2.98	.219
Anonymity*Singularity of Target	-0.13	0.06	3.55	.060
Anonymity*Privacy of Message	0.09	0.09	1.06	.497

Table 18

Fixed Effects Summary Table for Perceptions of the Person as Honest Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.54	0.09	59.03	<.001
Need for Chaos	-0.01	0.08	0.06	.941
Anonymity	-0.14	0.03	10.46	<.001
Singularity of Target	-0.04	0.04	1.02	.325
Privacy of Message	0.20	0.05	11.25	<.001
Anonymity*Singularity of Target	0.18	0.07	6.01	.009
Anonymity*Privacy of Message	-0.19	0.10	6.22	.007

Table 19

Fixed Effects Summary Table for Perceptions of Attention Seeking Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.42	0.11	31.69	<.001
Need for Chaos	0.09	0.08	0.49	.851
Anonymity	0.03	0.03	0.96	.237
Singularity of Target	0.03	0.04	0.64	.425
Privacy of Message	0.05	0.05	1.25	.095
Anonymity*Singularity of Target	0.18	0.07	5.87	.013
Anonymity*Privacy of Message	0.16	0.10	3.98	.039

Table 20

Results from Simple Slopes Analyses of Significant Interactions Between Anonymity and Singularity of Target

	Anonymity		*Singularity of Target*	
Perception	*r*	*t*	*r*	*t*
Honesty of Post	-0.29	-2.08*	0.38	2.97*
Honesty of Person	-0.31	-2.68*	0.36	2.82*
Attention Seeking	0.26	2.01*	-0.27	-2.05*

*$p \leq .05$

Table 21

Results from Simple Slopes Analyses of Significant Interactions Between Anonymity and Privacy of Message

	Anonymity		*Privacy of Message*	
Perception	*r*	*t*	*r*	*t*
Honesty of Post	-0.33	-2.74*	0.42	3.15*
Honesty of Person	-0.28	-2.04*	0.44	3.26*
Attention Seeking	0.25	1.98*	-0.35	-2.78*

*$p \leq .05$

References

Anderson, J., Bresnahan, M., & Musatics, C. (2014). Combating weight-based cyberbullying on Facebook with the dissenter effect. *Cyberpsychology, Behavior, and Social Networking*, *17*(5), 281-286.

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*(3), 335-344.

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment*, *14*(1), 16-37.

Bagby, R. M., & Marshall, M. B. (2003). Positive impression management and its influence on the Revised NEO Personality Inventory: a comparison of analog and differential prevalence group designs. *Psychological Assessment*, *15*(3), 333-342.

Bagby, R. M., Nicholson, R. A., Buis, T., Radovanovic, H., & Fidler, B. J. (1999). Defensive responding on the MMPI-2 in family custody and access evaluations. *Psychological Assessment*, *11*(1), 24-32.

Bargh, J. A., McKenna, K. Y., & Fitzsimons, G. M. (2002). Can you see the real me? Activation and expression of the "true self" on the Internet. *Journal of Social Issues*, *58*(1), 33-48.

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology*, *23*(1), 37-51.

Ben-Porath, Y. S., & Waller, N. G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment*, *4*(1), 14-26.

Bergstrom, K. (2011). "Don't feed the troll": Shutting down debate about community expectations on Reddit. com. *First Monday*, *16*(8).

Bianchi, E. C., Hall, E. V., & Lee, S. (2018). Reexamining the Link Between Economic Downturns and Racial Antipathy: Evidence That Prejudice Against Blacks Rises During Recessions. *Psychological Science*, *29*(10), 1584-1597.

Bourdieu, P. (2001). *Masculine domination*. Stanford University Press.

Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied Social Psychology*, *23*(23), 1933-1967.

Calvete, E., Orue, I., Estévez, A., Villardón, L., & Padilla, P. (2010). Cyberbullying in adolescents: Modalities and aggressors' profile. *Computers in Human Behavior*, *26*(5), 1128-1135.

Carr, C. T., Vitak, J., & McLaughlin, C. (2013). Strength of social cues in online impression formation: Expanding SIDE research. *Communication Research*, *40*(2), 261-281.

Cashel, M. L., Rogers, R., Sewell, K., & Martin-Cannici, C. (1995). The Personality Assessment Inventory (PAI) and the detection of defensiveness. *Assessment*, *2*(4), 333-342.

Cassidy, W., Jackson, M., & Brown, K. N. (2009). Sticks and stones can break my bones, but how can pixels hurt me? Students' experiences with cyber-bullying. *School Psychology International*, *30*(4), 383-402.

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, *65*(1), 57-70.

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:"On the Internet, Nobody Knows You're a Dog". *Computers in Human Behavior*, *23*(6), 3038-3056.

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continuum*, *23*(4), 525-535.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349-361.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784-793.

Dahlberg, L. (2001). The Internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society*, *4*(4), 615-633.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5-18.

Dilmac, B. (2009). Psychological needs as a predictor of cyber bullying: A preliminary report on college students. *Educational Sciences: Theory and Practice*, *9*(3), 1307-1325.

Donath, J. (1998). Identity and deception in the virtual community. *Communities in Cyberspace* ed. By Marc Smith and Peter Kollock. Routledge.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62-70.

Dyer, R., Green, R., Pitts, M., & Millward, G. (1995). What's the Flaming Problem? or Computer Mediated Communication-Deindividuating or Disinhibiting?. In *BCS HCI*(pp. 289-302).

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, *2*, 357-411.

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology*, *9*(1), 46-55.

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of White racism, arousal, and helping behavior. *Journal of Personality and Social Psychology*, *35*(10), 691-702.

Gaertner, S. L., & Dovidio, J. F. (1981). Racism among the well-intentioned. *Pluralism, racism, and public policy: The search for equality*, 208-222.

Gaertner, S. L., & Dovidio, J. F. (1986). *The aversive form of racism*. San Diego, CA, US: Academic Press.

Görzig, A., & Ólafsson, K. (2013). What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries. *Journal of Children and Media*, *7*(1), 9-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1476.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politenss Research, 6(2),* 215-242.

Harman, J. P., Hansen, C. E., Cochran, M. E., & Lindsey, C. R. (2005). Liar, liar: Internet faking but not frequency of use affects social skills, self-esteem, social anxiety, and aggression. *CyberPsychology & Behavior*, *8*(1), 1-6.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254-266.

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, *29*(2), 129-156.

Huang, Y. Y., & Chou, C. (2010). An analysis of multiple factors of cyberbullying among junior high school students in Taiwan. *Computers in Human Behavior*, *26*(6), 1581-1590.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*(2), 177-192.

Joinson, A. N. (2007). Disinhibition and the Internet. In *Psychology and the Internet (Second Edition)* (pp. 75-92). Academic Press. Alberta.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, *39*(10), 1123-1130.

Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *Journal of Adolescent Health*, *41*(6), S22-S30.

Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: a meta-analytic review. *Psychological Bulletin*, *130*(1), 143-172.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434-443.

Lea, M., & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing and Electronic Commerce*, *2*(3-4), 321-341.

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., & Rainie, L. (2011). Teens, Kindness and Cruelty on Social Network Sites: How American Teens Navigate the New World of" Digital Citizenship". *Pew Internet & American Life Project*.

Li, T. B. Q. (2005). Cyber-harassment: A study of a new method for an old behavior. *Journal of Educational Computing Research*, *32*(3), 265-277.

Locey, M. L., & Rachlin, H. (2015). Altruism and anonymity: A behavioral analysis. *Behavioural Processes*, *118*, 71-75.

Mandalaywala, T. M., Amodio, D. M., & Rhodes, M. (2018). Essentialism promotes racial prejudice by increasing endorsement of social hierarchies. *Social Psychological and Personality Science*, *9*(4), 461-469.

Martin, J. K., Pescosolido, B. A., & Tuch, S. A. (2000). Of fear and loathing: The role of'disturbing behavior,'labels, and causal attributions in shaping public attitudes toward people with mental illness. *Journal of Health and Social Behavior*, 208-223.

Miller, S. S., & Saucier, D. A. (2018). Individual differences in the propensity to make attributions to prejudice. *Group Processes & Intergroup Relations*, *21*(2), 280-301.

Miller, S. S., Martens, A. L., & Saucier, D. A. (2017). Attributions to Prejudice: Collective Anger and Action. *Understanding Angry Groups: Multidisciplinary Perspectives on Their Motivations and Effects on Society*, 29-40.

Moor, P. J., Heuvelman, A., & Verleur, R. (2010). Flaming on youtube. *Computers in Human Behavior*, *26*(6), 1536-1546.

Moore, M. J., Nakano, T., Enomoto, A., & Suda, T. (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, *28*(3), 861-867.

Moral-Toranzo, F., Canto-Ortiz, J., & Gómez-Jacinto, L. (2007). Anonymity effects in computer-mediated communication in the case of minority influence. *Computers in Human Behavior*, *23*(3), 1660-1674.

Nichols, D. S., & Greene, R. L. (1997). Dimensions of deception in personality assessment: The example of the MMPI-2. *Journal of Personality Assessment*, *68*(2), 251-266.

Nogami, T., & Takai, J. (2008). Effects of anonymity on antisocial behavior committed by individuals. *Psychological Reports*, *102*(1), 119-130.

Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, *4*(2), 148-169.

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*(6), 1137-1151.

Pauls, C. A., & Crost, N. W. (2005). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences*, *26*(4), 194-206.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811.

Postmes, T., & Baym, N. (2005). Intergroup dimensions of the Internet. *Intergroup communication: Multiple Perspectives*, *2*, 213-240.

Postmes, T., & Spears, R. (2002). Behavior online: Does anonymous computer communication reduce gender inequality?. *Personality and Social Psychology Bulletin*, *28*(8), 1073-1083.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research*, *25*(6), 689-715.

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, *27*(10), 1243-1254.

Price, M., & Dalgleish, J. (2010). Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. *Youth Studies Australia*, *29*(2), 51-63.

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., & Oksanen, A. (2016). Targets of online hate: Examining determinants of victimization among young Finnish Facebook users. *Violence and Victims*, *31*(4), 708-718.

Reader, B. (2012). Free press vs. free speech? The rhetoric of "civility" in regard to anonymous online comments. *Journalism & Mass Communication Quarterly, 89*(3), 495-513.

Reeckman, B., & Cannard, L. (2009). Cyberbullying: a TAFE perspective. *Youth Studies Australia, 28*(2), 41-54.

Retzlaff, P., Sheehan, E., & Fiel, A. (1991). MCMI-II report style and bias: Profile and validity scales analyses. *Journal of Personality Assessment*, *56*(3), 466-477.

Runions, K. C., (2015). Online moral disengagement, cyberbullying, and cyber-aggression. *Cyberpsychology, Behavior, and Social Networking*, *18*(7), 400-405.

Scandell, D. J., & Wlazelek, B. G. (1996). Self-presentation strategies on the NEO-Five Factor Inventory: Implications for detecting faking. *Psychological Reports*, *79*(3_suppl), 1115-1121.

Scott, C. R. (1998). The impact of physical and discursive anonymity on group members' multiple identifications during computer-supported decision making. *Western Journal of Communication (includes Communication Reports)*, *63*(4), 456-487.

Slonje, R., Smith, P. K., & Frisén, A. (2012). Processes of cyberbullying, and feelings of remorse by bullies: A pilot study. *European Journal of Developmental Psychology*, *9*(2), 244-259.

Son Hing, L. S., Chung-Yan, G. A., Grunfeld, R., Robichaud, L. K., & Zanna, M. P. (2005). Exploring the discrepancy between implicit and explicit prejudice: A test of aversive racism theory. In *Biennial meeting of the Society for the Psychological Study of Social Issues., Jun, 2002, Toronto, ON, Canada. March 2003, Sydney, Australia.*. Cambridge University Press.

Sticca, F., & Perren, S. (2012). Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*, *42*(5), 739-750.

Stratmoen, E., Lawless, T. J., & Saucier, D. A. (2019). Taking a knee: Perceptions of NFL player protests during the National Anthem. *Personality and Individual Differences*, *137*, 204-213.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321-326.

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human Communication Research*, *28*(3), 317-348.

Tynes, B., Reynolds, L., & Greenfield, P. M. (2004). Adolescence, race, and ethnicity on the Internet: A comparison of discourse in monitored vs. unmonitored chat rooms. *Journal of Applied Developmental Psychology*, *25*(6), 667-684.

Udris, R. (2014). Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior*, *41*, 253-261.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197-210.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197-210.

Walther, J. B., & Bazarova, N. N. (2007). Misattribution in virtual groups: The effects of member distribution on self-serving bias and partner blame. *Human Communication Research*, *33*(1), 1-26.

Waltman, M. S. (2018). The normalizing of hate speech and how communication educators should respond. *Communication Education, 67(2)*, 259-265.

Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., & Zhao, B. Y. (2014, November). Whispers in the dark: analysis of an anonymous social network. In *Proceedings of the 2014 Conference on Internet Measurement Conference* (pp. 137-150). ACM.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press.

Williams, D., & Skoric, M. (2005). Internet fantasy violence: A test of aggression in an online game. *Communication Monographs*, *72*(2), 217-233.

Wong-Lo, M., & Bullock, L. M. (2011). Digital aggression: Cyberworld meets school bullies. *Preventing School Failure: Alternative Education for Children and Youth*, *55*(2), 64-70.

Wright, M. F. (2013). The relationship between young adults' beliefs about anonymity and subsequent cyber aggression. *Cyberpsychology, Behavior, and Social Networking*, *16*(12), 858-862.

Wright, M. F., Harper, B. D., & Wachs, S. (2018). The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and Individual Differences*.

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry*, *45*(7), 1308-1316.

Appendix A - Demographics Materials

Demographics

Age: _____ **Ethnicity**: _____ **Hometown, State**: _____

Please indicate your biological sex (please circle one): *Male* *Female* *Other*: _____

Please indicate your gender identity (please circle one): *Man* *Woman* *Other*: _____

Please describe your hometown (please circle one): *Rural* *Suburban* *Urban*

Sexual Orientation (please circle one): *Heterosexual* *Homosexual* *Bisexual* *Other*:_____

Although it is often difficult to summarize one's political, economic, and social views in a single word or phrase, please indicate which of the following positions best represents your viewpoint (circle one):

a. *Liberal* b. *Middle of the road* c. *Conservative*

Please use the 9-point scale below to report your positions on the following three issues.

1 2 3 4 5 6 7 8 9

Very Liberal **Very Conservative**

1 _____ Foreign policy issues.

2 _____ Economic issues.

3 _____ Social issues.

Appendix B - Templates of Posts

Bailey [redacted]

14 hrs

Passed four Black guys on my way home and I've never been so scared in my life.

Example of a mock Facebook post

[–] strawberrylime 105 points 7 hours ago

My assigned partner in my Chem lab is this Black guy so I guess I'm screwed

permalink parent report reply

Example of a mock Reddit post

Appendix C - Examples of Hate Speech Items

Targeting Black People Generally:

1. Passed four Black guys on my way home and I've never been so scared in my life.
2. Black people whine and complain about being "oppressed" yet sit at home and collect welfare. It's called hard work!
3. If it weren't for football & basketball, Black men would literally have nothing.

Targeting Specific Black People:

1. Marc doesn't deserve to be on the basketball team. He's just there because he's Black
2. Keisha won't shut up about how Black people aren't as privileged as White people. Learn to work harder!
3. My assigned Chem lab partner is this guy named Deshawn. He's Black, so I guess I'm screwed.

Appendix D - Measures of Perceptions

1. This post is racist.
2. This post is meant to harm.
3. This post is meant to be hurtful.
4. This post is intended to be cruel.
5. The person who created this post is racist.
6. The person who created this post is cruel.
7. The person who created this post is hurtful.
8. The person who created this post meant to harm.
9. This post is meant to be taken honestly.
10. This post reflects the true beliefs of the person who made it.
11. The person who created this post believes that this post reflects reality.
12. The person who created this post wants others to believe that this post reflects reality.
13. The person who created this post just wants attention.